Descriptive Statistics

David A. Hughes, Ph.D.

Associate Professor
Auburn University at Montgomery
david.hughes@aum.edu

Once you have your data. . .

- Before we use our data to test hypotheses, it's best that we first summarize and describe our variables. (Why?)
- What kinds of information points help us to describe data?
- If you gathered some variable of interest, what are the sorts of things you'd want to know about it?

Frequency and rates

- We often want to measure the **frequency** with which phenomena occur—like how many Americans have health insurance.
- We could simply count up a set of observations, but we typically prefer to standardize frequencies.
- A **proportion** is measured on a scale from 0 to 1: $p = \frac{x}{n}$.
- Here, x represents the frequency with which an observation occurs, and n represents the sample size—or the number of opportunities there were for x to occur.
- If we multiply a proportion by 100, then we have a **percentage**: $\pi = 100\left(\frac{x}{n}\right)$.

Changes or differences in rates

- Sometimes, we like to describe how rates change or differ—like how did health insurance coverage change pre- versus post-Affordable Care Act?
- First, we could talk about a **percentage change**: $\pi_c = 100\left(\frac{\pi_2 - \pi_1}{\pi_1}\right)$, where π_1 represents an initial rate, and π_2 represents a subsequent rate.
- We could also consider a **percentage point change**: $\pi_{pc} = \pi_2 - \pi_1$.
- Example: In 2000, 86% of Americans had health insurance. In 2020, 92% did. What's the percent change and percentage point change in coverage?

Visualizing frequencies

- Often, when we want to visualize frequences, it's worth knowing whether a variable of interest is measured **categorically** (e.g., race or gender) or **continuously** (e.g., income or weight).
- **Histograms** and **density plots** are well-suited to depict frequencies for continuous-level data. **Pie charts** can be used for categorical data.
- We're also frequently interested in depicting how frequences change over time or across groups. Here, **bar charts** and **choropleth maps** can be quite useful as well.

Central tendency

- One of the first things we like to know when looking at a new variable is, "What's the average observation?"
- What we're really asking here is what the **central tendency** of a variable is.
- Most people assume the word, "average" refers to an **arithmetic mean**, but "average" could refer to any measure of central tendency.
- Some texts will refer to an **expected outcome** for a given variable, which is another way of discussing "averageness."

The mean

- The most widely used measure of central tendency is the arithmetic mean:
$$\bar{x} = \sum_{i=1}^{n} x_i \frac{1}{n}. \tag{1}$$
- Here, we simply add up all the observations in x and divide by the sample size, n.
- Note that if x is measured dichotomously with 0s and 1s, then Equation (1) simply gives a proportion.
- Means are appropriate for continuous-level variables.
- One downside to using means, however, is that they can be sensitive to outliers.

The median

- If we're worried about skewed data, the **median** could be a useful concept of averageness.
- The median of a variable arrayed from low-to-high is its middle-most observation, or the fiftieth percentile.
- If a variable has an even number of observations, Stata will take the two middle-most observations and calculate their mean. Nevertheless, both observations can be considered median.
- Medians can appropriately be used for continuous, ordinal, and dichotomous-level variables.

The mode

- The **mode** simply reports the most frequent observation in a variable.
- Arguably, modes are best suited for nominal-level data, but we could use them for any level of measurement.
- Note that, given a single-peaked, symmetric distribution for some variable, the mean, median, and mode will all be the same value.

Dispersion

- Some expected outcomes occur more frequently than others.
- For example, if I roll a pair of six-sided die, the mean outcome will be 7, and there's a 17% chance of rolling a 7.
- If I roll two twenty-sided die, the mean is now 11, but because I added more sides, I now only have a 10% chance of observing that mean value.
- This example illustrates a feature of variables known as **dispersion**. Some variables are fairly predictable while others are noisy.

Variance and standard deviation

- Probably the most widely used concepts to measure variable dispersion are **variance** and **standard deviaion**.
- Variance examines how far a given observation falls from its mean:

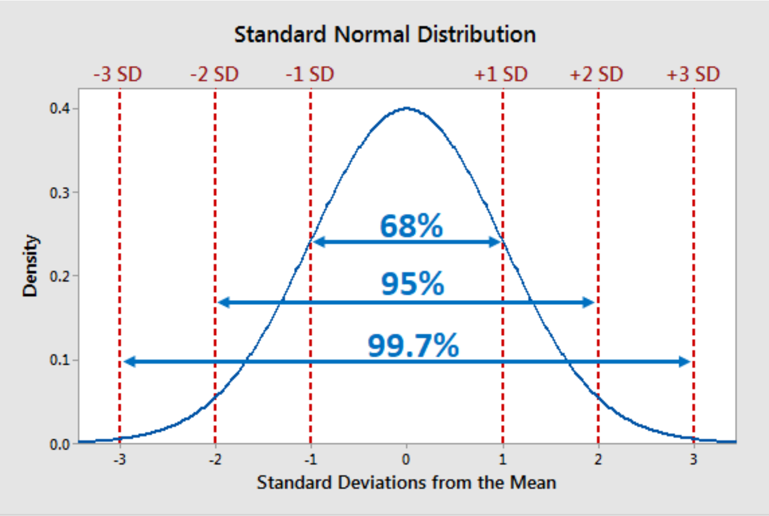
$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \tag{2}$$

- Standard deviation merely takes the square root of the variance:

$$\sigma = \sqrt{\sigma^2}. \tag{3}$$

- Variance and standard deviation are measured on a scale from 0 to $+\infty$, where greater values denote greater dispersion.
- We can interpret variance in units of x. For standard deviations, we use the **empirical rule**.

The empirical rule



Other measures of dispersion

- We might want to know other facts about a variable's dispersion aside from average distance to the mean.
- The **range**, for example, gives the minimum and maximum values for a variable.
- And the **interquartile range** gives the values of a variable at its 25th and 75th percentiles.
- Finally, it may be worth identifying **outliers** in a variable—a concept we'll return to later this semester.

Visualizing central tendency and dispersion

- We've already discussed how density curves and histograms can help illustrate frequency, but they can be useful as well for showing central tendency and dispersion.
- **Boxplots** are like density curves but smushed down flat. These are really useful for showing distributions across different categories of outcomes.
- We might want to observe changes in means over time, in which case **trend lines** can be useful, along with bar charts that include measures of variance.

Conclusion

- Before we can go about analyzing the relationships between variables, it is critical that we understand how every variable is measured.
- This not only helps us to make sure we don't make later mistakes but also helps our readers understand what typical observations look like.
- Presenting this information as clearly as possible only aids in this understanding.